SIGNAL PROCESSING V
THEORIES AND APPLICATIONS

Proceedings of EUSIPCO-90
Fifth European Signal Processing Conference
Barcelona, Spain, September 18–21, 1990

Edited by

Luis TORRES
Enrique MASGRAU
Miguel A. LAGUNAS

Department of Signal Theory and Communications
ETSIT-UPC
Barcelona, Spain

VOLUME II

1990

ELSEVIER
AMSTERDAM • NEW YORK • OXFORD • TOKYO

SIGNAL PROCESSING V: Theories and Applications
L. Torres, E. Masgrau, and M.A. Lagunas (eds.)
© Elsevier Science Publishers B.V., 1990

Adapting a Large Vocabulary Speech Recognition System to Different Tasks

P. Alto, M. Brandetti, M. Ferretti, G. Maltese, F. Mancini, A. Mazza, S. Scarci, G. Vitillaro

IBM Italy Rome Scientific Center via Giorgione 159, 00147 ROME (Italy)

The probabilistic approach to speech recognition has allowed the development of large-vocabulary, high-performances, real-time speech recognizers. Following this approach a speech recognition prototype for the Italian language has been built at the IBM Italy Rome Scientific Center. Many laboratory tests have shown the effectiveness of the prototype as a tool to create texts by voice. To assess the behavior of the recognizer in real environments it is necessary to adapt the vocabulary of the recognizer to different applications. In this paper we present the techniques needed to adapt the acoustic model and the language model, the results obtained for two different applications are also reported.

1. INTRODUCTION

In the last years the probabilistic approach to speech recognition has allowed the development of high-performances large-vocabulary speech recognition systems. At the IBM Rome Scientific Center a speech-recognition prototype for the Italian language, based on this approach, has been built. The prototype is able to recognize in real time natural-language sentences built using a vocabulary containing up to 20000 words. [1]. Once and for all the user has to perform an acoustic training phase (about 20 minutes long), during which he is required to utter a predefined text. Words must be uttered inserting small pauses (a few centiseconds), between them. The prototype architecture is based on a personal computer equipped with special hardware. The first system we developed was aimed at a business and finance lexicon. In the following we will refer to it as EF. This system was used to perform in-house experiments to assess the acceptance of the recognizer as a tool to create texts. These experiments showed the effectiveness of the prototype [2]. After this phase the necessity arose to perform experiments in real work environments. Two different applications were considered: the dictation of radiological reports and of insurance company documents. They will be indicated as RR and IR respectively. Due to their characteristics, these applications seemed to be very well suited for our purposes. To develop the systems to be employed during the experiments, we had to adapt the EF recognizer to the lexicon required by the new applications. In the probabilistic approach the vocabulary of the recognizer is predefined and no efficient way to adapt the vocabulary of the system exists. The paper describes the techniques we have used to solve the problem of vocabulary adaptation. The results obtained experimenting automatic text dictation during real work are also presented.

2. SYSTEM STRUCTURE

We look for the sequence of words $\hat{W}$ which has the highest probability given the acoustic information $\overline{A}$ extracted from the observed signal [3]. In our case the acoustic signal is a sequence of acoustic labels extracted from the signal every centisecond and representing the energy content of the signal in 20 frequency bands.
Applying the Bayes theorem we can write:

$$P(\overline{W}\,|\,\overline{A}) = \frac{P(\overline{A}\,|\,\overline{W})P(\overline{W})}{P(\overline{A})} \qquad (1)$$

where $P(\overline{W}\,|\,\overline{A})$ is the probability that the sequence of words $\overline{W}$ will produce the sequence of acoustic information $\overline{A}$. $P(\overline{W})$ is the *a priori* probability of the sequence of words $\overline{W}$. $P(\overline{A})$ is the probability of the sequence of acoustic information $\overline{A}$. We want to find the maximum of the above expression with respect to $\overline{W}$. We can ignore $P(\overline{A})$ because it does not depend on $\overline{W}$. Therefore we need to maximize the numerator of the expression (1).
The problem can be reduced to the following steps:

1. perform the signal processing stage to extract the acoustic information $\overline{A}$ from the speech signal;
2. compute the acoustic probability $P(\overline{A}\,|\,\overline{W})$ (this task is accomplished by the acoustic model);
3. compute $P(\overline{W})$ (this is done by the language model);
4. look for the most probable sequence of word through an efficient search strategy.

While the signal processing stage and the search strategy can be considered independent of the application, the acoustic and the language model must be changed according to the lexicon of the application. In the next paragraphs the techniques employed to adapt both models will be explained.

3. ACOUSTIC MODEL ADAPTATION

The acoustic model task is to compute $P(\overline{A}\,|\,\overline{W})$. In the probabilistic approach the acoustic model is based on hidden Markov models. An hidden Markov model is a finite state automata. For every time slice the model takes a transition from the current state to one of the allowed states (the transition can also produce no state changement). For each transition an acoustic label is produced [3]. Both the transitions and the label emission occur according two probability distributions. The distributions depend on the current state only. These models are called *hidden* because it is only possible to observe the sequence of acoustic symbols produced, while the sequence of states remains hidden.
Each word belonging to the vocabulary is represented by a different model.

Two different techniques exist to construct the models. The first one is based on the idea of automatically building the word model starting from several utterances of it produced by several speakers [4]. According to the second technique an alphabet of acoustic units to represent the basic sounds of the language is defined. The word model is built by concatenating the Markov models representing the acoustic units. In our case the latter technique was employed. Examples of acoustic units used for speech recognition are: syllables, diphones, phones. We choose the phone as phonetic unit. The basic sounds of the Italian language were described by a set of 56 phonetic units [1]. For each phonetic unit a Markov model representing its pronunciation has been created. In our system all the phonetic units have the same topological structure. The distinction between different sounds is left entirely to the probability distributions, called *model's parameters*. The computation of the parameters is accomplished during the acoustic training phase employing the predefined text uttered by the user.

According to the technique chosen, the first step that must be performed when adapting the recognizer to a new application, is to make the phonetic transcription of all the words in the vocabulary. To limit the number of the needed phonetic transcriptions, a database was built containing all the words and the phonetic transcriptions used in previous vocabularies. By using the database it is possible to find all the words for which a new phonetic transcription must be supplied.

Usually, the phonetic transcription is a performed manually. It is a very expensive process and for large vocabularies the transcriptions could contain errors. We tried to make the phonetic transcription process as automatic as possible. The systems that have been proposed to solve the problem of automatic phonetic transcription are based on rules [5] [6] or on automatic learning from training data [7]. Actually, these systems cannot provide the accuracy required for automatic speech recognition. This is due both to the complexity of the problem and to the difficulty to describe all the possibilities with a limited set of rules. We employed a different technique from the mentioned ones. We separated phonotactical knowledge (well described by a limited set of rules) from lexical knowledge (based on experience and not suitable for a formal description). Given the string representing the orthographic form of the word our system produces a set of phonetic transcriptions for that word, which are the ones that can be obtained applying our set of rules for the grapheme-to-phoneme translation. The user can choose manually the correct transcription on the basis of his lexical knowledge. Grapheme-to-phoneme translation for the Italian language has a relatively low uncertainty. A set of 78 rules allows to describe all the ambiguities. Each rule consists of a left part and a right part. The left part consists of a grapheme string and its (possibly empty) left and right graphemic contexts; the right part consists of the set of possible phonetic transcriptions for the grapheme string. The set of transcriptions produced applying this set of rules is then pruned by means of a set of global rules (which, for example, reject all the transcriptions which do not have one and just one stressed vowel). The right phonetic transcription always belongs to the resulting set. The average number of phonetic transcriptions per word is 5. Using this method it was possible to adapt rapidly the recognizer to the new application. The quality of the produced transcriptions was at least equal to a completely manual phonetic transcription.

4. IN-HOUSE TEST FOR THE NEW APPLICATION

Before experimenting the speech-recognizer in a real environment we needed to perform in-laboratory tests to assess the recognition rate of the system when used to dictate pre-defined texts. To perform the experiment a text containing phrases peculiar to the application must be created; it must be dictated by several different speakers. To make a meaningful test it is important that the text contains all the phonetic units used to build the acoustic model in phonetic contexts typical of the application.

Usually, the text is built manually trying to represent a large number of different contexts using the smallest number of sentences. To avoid this manual process a procedure has been built to prepare the test automatically. A set of sentences peculiar to the application is used as initial data. Usually the corpus employed to train the language model is used for this purpose. A preliminary analysis is done to eliminate all the sentences that contain words not included in the vocabulary. The first selected sentence is the one containing the greatest number of distinct phones. The sentences are then added incrementally, and at each step the sentence with the highest score among the selected ones, is chosen. The score is computed in the following way:

- the frequency of each phonetic unit in the previously selected sentences is computed;
- for each sentence in the available data and not yet selected a score is computed according to the following formula:

$$C(S_k) = \sum_i f_i \exp(-h_i) \qquad (2)$$

where f_i is the frequency of the phone i in the sentence S_k while h_i is the frequency of phone i in the sentences selected so far;

- the summation is extended only to the phones the frequency of which is less than a predefined threshold.

By applying this algorithm it is possible to select efficiently a set of sentences containing phonetic contexts typical for the application and suitable to assess the accuracy of the recognizer.

5. VOCABULARY ADAPTATION

In our system the vocabulary is predefined; this means that one of the most important factors affecting the usability of the speech recognizer is the availability of the largest number of words needed by the user to create the text. The choice of the vocabulary is a key factor for the system performances.

In our first experiment a vocabulary containing more than 20000 words was used. This vocabulary (EF) is aimed at the dictation of economy and finance reports. The 20000 words were chosen as the most frequent in a corpus containing 44 millions of words composed by: articles from the most important Italian economy and finance newspaper (*Il Sole 24 Ore*), articles from an economy and finance newsmagazine (*Il Mondo*), and press agency news. The coverage of this vocabulary computed on a disjoint corpus was 96.5%

The economy and finance lexicon is very different from the lexicon required to dictate radiological reports, while it has

some similarities with respect to the lexicon used in the insurance company reports. In the first case (RR) the vocabulary was selected by using a corpus containing only radiological reports; in the second case (IR) the vocabulary was built taking EF as a starting point.

Radiological Reports Vocabulary

The available corpus contained about 5 million words (we will call this corpus *HO*) collected in four different hospitals (*HO*1, *HO*2 , *HO*3, *HO*4). The hospital were the experiment was held provided us with a corpus containing only 50000 words (*HE*). The first problem was the choice of the vocabulary size. We adopted the criterion of analyzing the variation of the coverage with respect to the vocabulary size (figure 1).

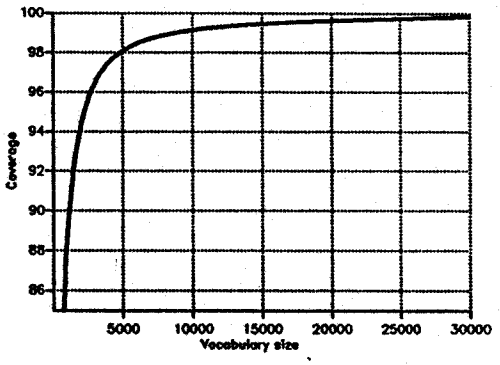

Figure 1. Coverage of corpus HO as function of vocabulary size.

A vocabulary containing 5000 words seemed to us a reasonable trade-off between the need to reach high coverage and the need to have enough data to estimate the language model parameters.

One of the main characteristics found in this kind of lexicon was the presence of a set of words peculiar to the report dictation process at each location. To make the recognizer well suited to the needs of the experimenter all the 3200 different words found in corpus *HE* were included in the vocabulary. The vocabulary was completed adding 1900 words which were the most frequent ones in *HO* corpus not included in the previous list of 3200 words. The *HO* words were ordered according to the average frequency of each word in the various corpora HO_i:

$$\bar{f} = \frac{1}{4} \times \sum_{i=1}^{4} \frac{C_{HOi}(w)}{N_{HOi}} \qquad (3)$$

$C_{HOi}(w)$ is the number of word w occurrences in corpus HO_i while N_{HOi} is the size of corpus HO_i). The number of occurrences was normalized because the four corpora HO_i have different sizes. The resulting vocabulary had a 100% coverage on reports from *HE* and 97.5% coverage on *HO* reports.

Insurance Company Reports Vocabulary

The selection of the words to be used in the IR vocabulary was performed by analyzing a corpus (*IC*) containing 1.5 million words provided by the insurance company. There was a certain similarity between the EF and IR lexicons: the coverage obtained by EF vocabulary on *IC* corpus was 95.2%. The 15000 most frequent words in EF were selected and the 3100 most frequent words found in *IC* not contained in the previous selected words were added. The resulting vocabulary IR showed a 99% coverage on *IC* texts and a 95.7% coverage on economy and finance texts.

6. LANGUAGE MODEL CONSTRUCTION

The task of language model is to compute $P(\overline{W})$. The computation is performed in the following way:

$$P(\overline{W}) = \prod_{i=1}^{N} P(w_i | w_{i-1}, w_{i-2}) \qquad (4)$$

This is the so-called trigram language model[8] : it contains the approximation of considering equivalent all the sentences ending with the same couple of words $w_{i-1}w_{i-2}$. The number of possible trigrams is so large that is practically impossible to collect the amount of data needed to estimate the probability of each of them. To overcome the problem an interpolation between different probability distributions is performed. Three different distributions are computed for trigrams, bigrams and unigrams. The probability of word w_3 given the words w_1 and w_2 is estimated as follows:

$$P(w_3 | w_1 w_2) = \lambda_3 \frac{C(w_1 w_2 w_3)}{C(w_1 w_2)} + \lambda_2 \frac{C(w_2 w_3)}{C(w_2)} + \lambda_1 \frac{C(w_3)}{N} + \lambda_0 \frac{1}{V}$$

where $C(w_1, \ldots, w_n)$ is the number of times the word string $w_1, \ldots, w_n$ was observed in the training data and V is the number of words in the vocabulary. The λ coefficients are estimated using the *Expectation-maximization* algorithm [9]. The trigram language model is an effective tool to represent the linguistic constraints for speech recognition purposes; on the other hand it requires a large amount of training data. The larger is the training corpus, the higher is the recognition rate [10]. The models for the two vocabularies RR and IR were built following the previously described technique. The RRlanguage model was trained using 4.8 million words. In the IR case a corpus containing 1.5 million words typical of the application was merged with 40 million words extracted from the EF corpus. In the following table the most significant parameters of the language models are reported.

Table 1. Language models characteristics		
Parameter	RR	IR
Vocabulary size	5100	18100
Training data	4.8	1.5 + 40
Millions of different trigrams	0.62	3.4
Millions of different bigrams	0.19	2.3
Perplexity	38	18

Perplexity is a typical measure of the predictive power of a language model [11]. It estimates the average number of words that are considered equiprobable by the model.

7. RESULTS

In the following paragraph the results obtained experimenting the two prototypes during real work are reported.

Radiological Reports Dictation

Four doctors have used the recognizer during their every-day work to prepare the reports to be delivered to the patients. No one had any difficulty in inserting short pauses between words. The doctors dictated 150 reports containing more than 12000 words. The vocabulary coverage for the dictated reports was 98.6%. Table 2 reports the results of the experiment.

Table 2. Radiological reports dictation			
Speaker	Number of reports	Error rate	Speaker's errors
sp1	25	1.7%	1.4%
sp2	14	2.0%	1.2%
sp3	76	3.5%	0.9%
sp4	35	5.0%	3.2%

The error rate is referred to the number of errors done by the recognizer without taking into account errors due to words not included in the vocabulary. The speaker's error rate is referred to the errors due to misuse of the recognizer (wrong commands, absence of pause between words, etc.). The global error rate can be computed summing the numbers in the third and fourth column of the table. We can see that the recognizer's performances ranges from 90.5% to 95.6% of accuracy.

Insurance Company Reports Dictation

The experimentation was carried on by five different users who dictated more than 8000 words. The vocabulary coverage on the dictated text was about 99%. Table 2 reports the results obtained in this case.

Table 3. Insurance company reports dictation		
Speaker	Error rate	Speaker's error rate
sp1	1.9%	1.0%
sp2	1.4%	0.5%
sp3	14.0%	2.0%
sp4	7.4%	1.5%
sp5	2.4%	0.8%

REFERENCES

[1] P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, G. Volpi, **Large-Vocabulary Speech Recognition: a System for the Italian Language,** *IBM Journal of Research and Development,* Vol. 32, No. 2, March 1988, pp.217-226.

[2] P. Alto, M. Brandetti, M. Ferretti, G. Maltese, S. Scarci, **Experimenting Natural-Language Dictation with a 20000-Word Speech Recognizer,** *IEEE CompEuro 89,* Hamburg, May 8-12, 1989, pp. 2-78 - 2-81.

[3] L.R. Bahl, F. Jelinek, R.L. Mercer, **A Maximum Likelihood Approach to Continuous Speech Recognition,** *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. PAMI-5, no. 2, 1983, pp. 179-190.

[4] L.R. Bahl, P.F. Brown, P.V. De Souza, R.L. Mercer, M.A. Picheny, **Acoustic Markov Models Used in the Tangora Speech Recognition System,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,*

[5] R. Carlson, B. Granstroem, **A Text-to-Speech System Based Entirely on Rules,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,* Philadelphia, PA, April 1976.

[6] D. H. Klatt, **Structure of a Phonological Rule Component for Synthesis-by-Rule Program** *IEEE Trans. on Acoust., Speech and Sig. Proc.,* vol. ASSP-24, no. 5, 1976, pp. 391-398.

[7] T. J. Sejnowski, C. R. Rosenberg, **Parallel Networks that Learn to Pronounce English Text,** *Complex Systems,* 1 (1987), pp. 145-168.

[8] F. Jelinek, **The development of an experimental discrete dictation recognizer,** *Proceedings IEEE,* vol. 73, no. 11, November 1985, pp. 1616-1624.

[9] F. Jelinek, R.L. Mercer, **Interpolated Estimation of Markov Source Parameters from Sparse Data,** *in "Pattern Recognition in Practice", E.L.Gelsema and L.N. Kanal, Ed.,* North-Holland, New York, 1980, pp. 381-387.

[10] M. Ferretti, G. Maltese, S. Scarci, **Measures of Language Model and Acoustic Information in Probabilistic Speech Recognition,** *Eurospeech 89,* Paris, September 1989, pp. 473-476.

[11] F. Jelinek, R.L. Mercer, L.R. Bahl, J.K. Baker, **Perplexity - a Measure of Difficulty of Speech Recognition Tasks,** *94th Meeting Acoustical Society of America,* Miami Beach, December 1977.